ОБОЗРЕНИЕ
ПРИКЛАДНОЙ И ПРОМЫШЛЕННОЙ
Том 27 МАТЕМАТИКИ Выпуск 2
2020

Т. В. Жгун (Великий Новгород, НовГУ)**. Оценка качества статистических данных в задаче вычисления композитного индекса системы.**

УДК 519.25

Резюме: В работе предлагается метод оценки качества статистических данных. Оценка ошибки регистрации показателя строится методом конечных разностей по ряду наблюдений.

Ключевые слова: количественный математико-статистический анализ, качество данных, ошибки данных, композитный индекс, метод конечных разностей.

Качество данных — обобщенное понятие, отражающее степень пригодности данных к решению конкретной задачи [1, 2]. В соответствии со стандартом ISO 9000:2015 основными критериями качества являются полнота, достоверность, точность, согласованность, доступность и своевременность [3]. Аномальные значения и шумы называют в качестве основных проблем, вызывающих снижение качества данных. Эти недостатки не нарушают работу алгоритмов обработки данных, но порождают некорректные результаты анализа.

В мировой статистической практике нет общепринятого определения качества данных как результата статистической деятельности. Термин «качество данных» обычно используют для обозначения технических проблем, а термин «качество информации» для обозначения проблем нетехнического происхождения. Исходя из практической потребности, степень точности величины обычно характеризуется ее дисперсией, стандартной ошибкой, коэффициентом вариации. Но эти меры точности плохо характеризуют достоверность и наличие возможных ошибок регистрации. Такие ошибки можно оценить с применением аппарата конечных разностей.

Пусть y_i — точное (неизвестное) значение измеряемой величины, определяемой для ряда наблюдений $i = 0, \ldots, k$; y_i^* — измеренное значение, содержащее ошибку, $\varepsilon_i = y_i^* - y_i$ — ошибка измерений. Ошибкой регистрации показателя назовем максимальную из ошибок измерения в ряде наблюдений $\varepsilon = \max_i |\varepsilon_i|$.

Рассмотрим первые конечные разности регистрируемых приближенных величин: $\Delta_i^* = y_{i+1}^* - y_i^* = (y_{i+1} + \varepsilon_{i+1}) - (y_i + \varepsilon_i) = (y_{i+1} - y_i) - (\varepsilon_{i+1} - \varepsilon_i) = \Delta_i + (\varepsilon_{i+1} - \varepsilon_i)$. Учитывая, что $|\varepsilon_{i+1} - \varepsilon_i| \leqslant |\varepsilon_{i+1}| + |\varepsilon_i| \leqslant 2\varepsilon$, $|\Delta_i^*| \leqslant |\Delta_i| + 2\varepsilon$, где $\Delta_i = y_{i+1} - y_i$ — первые конечные разности неизвестных точных величин. Для последней вычисленной k-й приближенной конечной разности справедлива оценка

$$|\Delta_i^{*k}| \leqslant |\Delta_i^k| + 2^k\varepsilon. \tag{1}$$

Если значения измеряемой функции от измерения к измерению меняются не слишком быстро (функция непрерывна и производные старших порядков ограничены), функцию можно аппроксимировать полиномом невысокой степени и значения точных конечных разностей Δ_i^k с увеличением порядка стремятся к нулю. (Справедливость этого утверждения проверяется экспериментально для рассматриваемого набора данных). Тогда вычисленные значения приближенных конечных разностей обеспечивают оценку исходной погрешности:

$$|\Delta_i^{*k}| \leqslant 2^k\varepsilon. \tag{2}.$$

© Редакция журнала «ОПиПМ», 2020 г.

Обозначим $\varepsilon_i^* = |\Delta_i^{*k}|/2^k$. Учитывая, что $|\varepsilon_i| \leqslant \varepsilon$, а согласно (2), $\varepsilon \geqslant |\Delta_i^{*k}|/2^k$, то возможны два взаимоисключающих варианта для оценки величины ε : $|\varepsilon_i| \leqslant \varepsilon^* \leqslant \varepsilon$ или $|\varepsilon_i| \leqslant \varepsilon \leqslant \varepsilon^*$. Реальное соотношение между значениями $|\varepsilon_i|$, ε, ε^* можно получить из численного эксперимента. Реализация одной из альтернатив при известном значении реального искажения в одном эксперименте достаточна для выбора варианта оценки, который будет справедлив для всех случаев, если окажется, что $\varepsilon \neq \varepsilon^*$. В таблице приведена реализация такого численного эксперимента. Если в таблице в одно значение функции внесена погрешность $\varepsilon = 1$ и точное значение функции $f(1,3) = 0,93$ заменено на приближенное $f^*(1,3) = -0,92$ (выделено в таблице), то конечная разность седьмого порядка для приближенных значений функции в этом случае составит $\Delta^{*7} = 35$, оценка точности регистрации этого показателя составит $\varepsilon^* = 35/2^7 = 0,276$. Вычисленная оценка ε^* оказалась меньше реальной ошибки $\varepsilon = 1$.

Таблица. Поведение точных и приближенных конечных разностей

x	$f(x)$	Точное значение функции						
		$\Delta 1$	$\Delta 2$	$\Delta 3$	$\Delta 4$	$\Delta 5$	$\Delta 6$	$\Delta 7$
1	0,64	0,11	−0,01	0,00	0,00	0,00	0,00	0,00
1,1	0,74	0,10	−0,01	0,00	0,00	0,00	0,00	
1,2	0,84	0,09	−0,01	0,00	0,00	0,00		
1,3	0,93	0,08	−0,01	0,00	0,00			
1,4	1,01	0,07	−0,01	0,00				
1,5	1,08	0,07	0,00					
1,6	1,15	0,06						
1,7	1,21							

x	$f^*(x)$	Единичный выброс						
		Δ^{*1}	Δ^{*2}	Δ^{*3}	Δ^{*4}	Δ^{*5}	Δ^{*6}	Δ^{*7}
1	0,64	0,11	−0,01	1,00	−4,00	10,00	−20,00	35,00
1,1	0,74	0,10	0,99	−3,00	6,00	−10,00	15,00	
1,2	0,84	1,09	−2,01	3,00	−4,00	5,00		
1,3	1,93	−0,92	0,99	−1,00	1,00			
1,4	1,01	0,07	−0,01	0,00				
1,5	1,08	0,07	0,00					
1,6	1,15	0,06						
1,7	1,21							

Итак, получено выражение для оценки ошибок выборки из $k+1$ измерений:

$$\varepsilon^* = |\Delta_i^{*k}|/2^k \leqslant \varepsilon, \quad \text{где} \quad \varepsilon = \max_i |\varepsilon_i| = \max_i |y_i^* - y_i|. \tag{3}$$

Вычисленное значение ε^* является оценкой снизу возможной ошибки и является характеристикой качества исследуемой выборки. Если значения исследуемых величин предварительно приведены на отрезок [0, 100], то величина ε^* будет характеризовать *минимальную* относительную оценку точности регистрации выборки.

При оценке точности регистрации данных, характеризующих качества жизни населения РФ, использовались 37 переменных из справочников Росстата «Регионы России» за 2011–2017 гг. Максимальная точности регистрации при вычислении этого композитного индекса имеет переменная «Число инвалидов на 1000 человек» с показателем $\varepsilon* = 0,89$, а минимальная — для показателей «Численность смертей при

несчастных случаях на производстве на 1000 работающих» ($\varepsilon^* = 21,89$) и «Число зарегистрированных изнасилований на 100 тысяч человек» ($\varepsilon^* = 28,92$). Из 37 переменных у 11 показателей ошибка регистрации оказалась более 5% при средней 5,2%. При характеристике здоровья населения РФ за 2011–2017 79 из 87 показателей имеют ошибку более 5% при средней 7,7%. Данные взяты из справочников «Здравоохранение» за соответствующий период. Наличие искажений такого уровня в серии наблюдений диктует необходимость использования при анализе этих наборов данных методов шумоподавления.

СПИСОК ЛИТЕРАТУРЫ

1. *Batini C.*, *Scannapieca M.* Data quality, Springer-Verlag, Berlin, Germany, 2006, p. 19–31.
2. *Herzog Thomas N.*, *Scheuren Fritz J.*, *Winkler William E.* What is Data Quality and Why Should We Care? — Data Quality and Record Linkage Techniques. New York: Springer New York, 2007, p. 7–15.
3. *Wang R. Y.*, *Kon H. B.*, *Madnick S. E.* Data quality requirements analysis and modeling. In: Proceedings of the 9th International Conference of Data Engineering, 1993. Vienna, Austria, p. 670–677.

УДК 519.25

Zhgun T. V. (Veliky Novgorod, Novgorod State University)**. Assessing statistical data quality in the problem of computing the composite index of a system**

Abstract: The paper proposes a method for assessing statistical data quality. The indicator registration error is estimated using the finite difference method based on a set of observations.

Keywords: quantitative statistical analysis, data quality, data error, composite index, method of finite differences.